Research Report

Severity of Organized Item Theft in Computerized Adaptive Testing: An Empirical Study

Qing Yi

Jinming Zhang

Hua-Hua Chang

Research & Development

July 2006
RR-06-22

Severity of Organized Item Theft in Computerized Adaptive Testing:
An Empirical Study

Qing Yi
Harcourt Assessment, Inc., San Antonio, TX

Jinming Zhang
ETS, Princeton, NJ

Hua-Hua Chang
University of Illinois at Urbana-Champaign

July 2006

Abstract

Chang and Zhang (2002, 2003) proposed several baseline criteria for assessing the severity of possible test security violations for computerized tests with high-stakes outcomes. However, these criteria were obtained from theoretical derivations that assumed uniformly randomized item selection. The current study investigated potential damage caused by organized item theft in computerized adaptive testing (CAT) for two more realistic item selection methods, the maximum item information and the *a*-stratified, while using the randomized method as a baseline for comparison. The results of the study indicated that the damage could be very severe, especially when the thieves took the test in the early stage of utilization of an item pool. Among the three CAT methods examined in this study, the maximum item information method with Sympson-Hetter exposure control was most vulnerable to organized item theft.

Key words: Test security, computerized adaptive testing, organized item theft, item selection methods

Acknowledgments

The authors are grateful to Dan Eignor, Tim Davey, and Frederic Robin for their helpful comments on an earlier version of the manuscript.

Introduction

Computerized adaptive testing (CAT) has the capability of administering a test to small groups of examinees at frequent adjacent time intervals, which is referred to as *continuous testing*. This provides examinees with the flexibility of scheduling a test. For example, the computerized Graduate Record Examinations® (GRE®) General Test is offered to examinees year-round in the United States, Canada, and many other countries, whereas the traditional paper-and-pencil (P&P) version of the same test was administered three times a year but only twice in some Asian countries. However, the weakness of CAT also lies with continuous testing because examinees who take a test earlier may share information with examinees who will take the test later, which increases the risk that many items may become known to examinees before they actually take the test. The activities of memorizing and sharing test information among examinees can inflate test scores of examinees who have gained preknowledge of the test while punishing honest examinees and hence threaten the validity of a test.

To reduce the impact of item sharing, item exposure rates should be controlled. The exposure rate of an item is defined as the ratio between the number of times the item is administered and the total number of examinees. A closely related index to the item exposure rate is the test overlap rate, which was originally defined as the average of the percentage of items shared by a pair of examinees across all such pairs (Way, 1998; Chen, Ankenmann, & Spray, 2003). Chang and Zhang (2002, 2003) generalized the definition of test overlap rate from the original two examinees to a group of m examinees. They also derived the theoretical distributions for item sharing and item pooling indices. According to Chang and Zhang, many rules currently employed in large-scale CAT programs were obtained from previous out-of-date empirical studies. For example, Stocking's rule of thumb (see Way, 1998) requires that an item pool size should be 12 times the test length. These rules may need to be modified and improved; moreover, new rules need to be developed.

Clearly, CAT test security issues must be studied in a broad context, and new emphasis should be placed on organized item theft. The objective of this paper is to empirically investigate how organized item theft could cause damage to CAT. Different CAT designs may yield different item exposure and test overlap rates, and our investigation focused on two item selection methods that have been researched extensively. One is the maximum information method (Lord, 1980) and the other is the *a*-stratified method (e.g., Chang & Ying, 1999; Chang, Qian, & Ying, 2001; Yi & Chang, 2003). The number of compromised items is also highly sensitive to the item selection method adopted in the CAT design. To establish a benchmark, an item selection algorithm that is based on purely randomized item selection was used as a baseline. The randomized item selection method equalizes item exposure rates and hence yields the best level of test security when compared with all other item selection methods. Our research interest was in assessing the severity of possible test security violations caused by organized item theft. More specifically, for a given CAT design, how many items could be compromised by forming an organized item theft group with differing numbers of thieves? For a collection of compromised items, what would be the possibility of each examinee encountering these items in his/her test? Since the process of examinees taking a CAT can be modeled as a time series, it is interesting to explore the time effect of sending thieves to take tests. Would it cause more severe damage to send thieves to take tests earlier in the life of an item pool than sending them later? In this study, the term *damage* had a broad sense, that is, the possibility of examinees encountering compromised items. Obviously, the use of compromised items by some examinees may lead to test score inflation. However, it may be too complicated to model cheating behavior as to who will use the compromised items and who will not. As an initial empirical investigation, we only focused on organized item theft; therefore, our simulation design was to randomly select a group of examinees as thieves who intentionally memorize test items.

This paper first briefly describes the CAT methods included in the study. This is followed by a section summarizing the results from Chang and Zhang’s (2002) theoretical derivations. The next section is on methodology and describes the details of the simulation procedure involved in the study. The last section contains concluding remarks and also discusses future research directions.

CAT Methods Investigated

One of the most commonly used item selection methods in CAT is based on maximum item information, which yields the best measurement efficiency; however, it does not include any item exposure control mechanism. Therefore, this method needs to be incorporated with an item exposure control procedure to achieve better test security management. Different methods of item exposure control have been proposed by various researchers (e.g., Davey & Parshall, 1995; Hetter & Sympson, 1997; Stocking & Lewis, 1998; Sympson & Hetter, 1985; Thomasson, 1995). The Sympson-Hetter (SH) procedure uses item exposure control parameters to probabilistically control the frequencies with which items are administered (Hetter & Sympson, 1997; Sympson & Hetter, 1985). SH exposure control parameters are obtained through a series of simulated CAT administered to a target population. After obtaining the exposure control parameters, they are used in a CAT design to control the frequency with which items are administered. The maximum item information method with the SH item exposure control (MII-SH), which incorporates SH into the maximum item information selection method, is one of the most widely used item selection methods for limiting the items' exposure rates to a prespecified value.

The *a*-stratified methods (Chang & Ying, 1999; Chang et al., 2001; Yi & Chang, 2003) select items from a stratified pool based on the closeness between item difficulty and the current CAT ability estimate. In the *a*-stratified method with content blocking (STRC; Yi & Chang, 2003), an item pool is first divided into groups based on the content specifications of the pool. Within each content group, the steps of the *a*-stratified with *b* blocking method as described in Chang et al. (2001) are followed to obtain several strata. The resulting stratified pool has the following three characteristics: (a) the content coverage of each stratum is similar to that of the whole item pool; (b) the distribution of *b*-parameters in each stratum is as similar as possible to that of the item pool; and (c) the average value of *a*-parameters increases across strata. The test is divided into several stages, one per stratum. STRC then selects items from the corresponding strata based on the match between item difficulty and an examinee's current CAT ability estimate. Items from the stratum with low average *a*-values are administered in the early stages of the test and items with high average *a*-values are used during the later stages. The SH exposure control procedure can also be incorporated into STRC (STRC-SH) to achieve the goal of limiting the maximum observed item exposure rate to a prespecified level.

The randomized item selection method, as indicated by its name, randomly selects items from the whole item pool. It results in roughly equalized item exposure rates; thus, there is no need to incorporate SH in this method.

Chang and Zhang's Theoretical Results

Chang and Zhang (2002) derived the *item sharing* and *item pooling* indices to compute the degree of possible test security violations based on the randomized item selection procedure. For α randomly sampled examinees, let X_α be the number of common items shared by these examinees. The item sharing index is then defined as the expected value of X_α, that is, $E[X_\alpha]$.

Item pooling, on the other hand, is when one examinee (beneficiary) gathers information from several examinees who have taken the test (nonbeneficiaries). Let Y_α be the number of items one examinee can obtain from α examinees. The expectation of Y_α, $E[Y_\alpha]$, is the item pooling index. With a randomized item selection procedure, Chang and Zhang derived the theoretical distributions for both X_α and Y_α for any given α.

Since test overlap rates are highly sensitive to methods used in item selection, ability estimation, and item exposure control, one must search for the most promising candidate from several possible CAT designs. Chang and Zhang's indices can serve as a benchmark for practitioners to evaluate a particular combination of testing settings. The discrepancy between the theoretical lower bounds and the observed rates obtained from a specific CAT design can provide information about the security prospects of this design. A large difference indicates the selected design needs to be improved to reduce the observed test overlap rate, while a small discrepancy demonstrates little improvement is needed.

Chang and Zhang (2003) extended their research in examining the issue by asking how many thieves are needed to compromise a certain proportion of an item pool. Their findings indicated that potential test security violations can be lessened if a large number of items are included in an item pool. For example, with an item pool of 1,000 items, if each thief can memorize 20 items, then 34 thieves are needed to compromise 50% of the item pool. However, if the item pool consists of 500 items, and each thief still can memorize 20 items, then only 17 thieves are needed to compromise 50% of the item pool.

Chang and Zhang's (2002, 2003) research focused on computerized tests using a randomized item selection procedure. Empirical research is needed to investigate the potential damage caused by organized item theft in CAT using more realistic CAT item selection methods. The findings of such research can provide guidance to practitioners in designing more secure CAT.

Simulation Design

Simulation studies were conducted to investigate the effects of applying different strategies in organized item theft in CAT when two item selection methods were used, the MII-SH and STRC-SH methods. The randomized item selection method was also used to serve as a baseline for comparison. The item pool consists of 480 multiple-choice items from a large-scale achievement test. There are three content areas in this test, in which 40% of the items are from content area one, and 30% of the items are from content areas two and three. The three-parameter logistic (3-PL) item response theory (IRT) model was assumed and the BILOG computer program (Mislevy & Bock, 1982) was used to calibrate the item parameters. The means of the calibrated *a*-, *b*-, and *c*-parameters are 1.056, 0.111, and 0.191, with standard deviations of 0.347, 1.060, and 0.085, respectively.

CAT Simulations

Ten thousand θ values were generated from a standard normal distribution. For each simulee, a fixed length CAT of 40 items was simulated. A content control procedure that uses a modified multinomial model as described in Yi and Chang (2003) was implemented as part of the CAT methods so that each simulated CAT consists of about 40% items of content area one and 30% items from each of the other two content areas.

Following the steps in Yi and Chang (2003), the item pool was stratified into four strata for the STRC-SH method. The first item was randomly selected from a list of 10 optimal items assuming an examinee's initial ability estimate of -1, without content balancing constraints. More specifically, for STRC-SH, 10 items were selected from the first stratum according to the closest match between item difficulty and the ability estimate of -1; for the MII-SH procedure, the 10 most informative items were selected at the ability estimate of -1; and for the randomized method, 10 items were randomly selected at the ability estimate of -1. The first item was then randomly selected from these 10 items.

The rest of the items from the designated content areas were selected based on the item selection criteria endorsed by each of the methods. For STRC-SH, the next item was selected if the following two conditions were satisfied: (a) the item had the closest match between item difficulty and the current CAT ability estimate; and (b) a uniform random number was less than or equal to the item exposure control parameter. For MII-SH, the next item was selected if the following two conditions were met: (a) the item had the maximum information at the current CAT ability estimate; and (b) a uniform random number was less than or equal to the item exposure control parameter. For the randomized procedure, the next item was randomly selected from the whole item pool. For both the STRC-SH and MII-SH procedures, item exposure control parameters were obtained through a series of simulated CATs administered to 10,000 simulees, and the maximum item exposure rate was set at 0.20. If the second condition for item selection listed previously was not met, then the next optimal item was selected and its exposure control parameter was compared to a new uniform random number.

The expected a posterior (EAP) method was used to estimate ability initially, until at least one correct and one incorrect item response were obtained, and five items had been administered. Afterwards, maximum likelihood estimation (MLE) was used.

Thieves and Compromised Items

Examinees who intentionally memorize items during testing and then share these items with other examinees who will take the test later were defined as thieves in this study. Based on the items administered to the 10,000 simulees, the thieves and items compromised by the thieves were randomly selected from the population and the items administered to these thieves, respectively. The numbers of thieves were set at 10, 20, and 30, while the number of items that a thief could memorize was 10. There were four ways to design the time when the thieves were actually taking the tests; the thieves randomly appeared in: (a) the 10,000 simulees; (b) the first 1,000 simulees; (c) the first 5,000 simulees; and (d) the last 5,000 simulees. This design takes consideration of the effect of time in assessing the severity of organized item theft occurred for different time sequences. In our simulation, we first randomly selected a thief, and then 10 out of the 40 items administered to the thief were randomly selected as the compromised items.

Evaluation Criteria

The number of items falling into various ranges of the observed item exposure rate (r) was summarized. The χ^2 index, a measure used to quantify the equalization of item exposure rates, was computed:

$$\chi^2 = \frac{\sum_{i=1}^{N} (r_i - L/N)^2}{L/N}, \tag{1}$$

and

$$r_i = \frac{\text{number of times the } i^{th} \text{ item is used}}{M}, \tag{2}$$

where N represents the size of an item pool, L denotes the length of a test, and M is the number of examinees. Note that L/N denotes a desirable uniform rate for all items, and Equation (2) represents the observed item exposure rate. The observed test overlap rate was computed as the average of the percent of the common items shared by a random pair of examinees across all such pairs.

Measurement precision is usually evaluated based on the difference between the estimated and true θ value. In this study, we computed bias and root mean square error (RMSE) to evaluate the measurement precision.

$$bias = \frac{\sum_{m=1}^{M}(\hat{\theta}_m - \theta_m)}{M}, \tag{3}$$

and

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta_m)^2}, \tag{4}$$

where M is the number of examinees, $\hat{\theta}_m$ is the estimated ability of examinee m ($m = 1, 2, \ldots, M$), and θ_m is the true ability of this examinee. The correlation coefficient between $\hat{\theta}_m$ and θ_m ($\rho_{\hat{\theta}_m, \theta_m}$) was calculated when evaluating the overall measurement precision. The conditional bias and RMSE were obtained at nine equally spaced θ points from –2 to 2 in increments of 0.5 with 10,000 replications at each of the θ points.

The total number of compromised items obtained from each condition and the average number of compromised items each examinee encountered were calculated. The damage caused by the compromised items can be quantified as the average number of compromised items each examinee can encounter after thieves having gathered information from taking the test. Based on the derivation of Zhang and Chang (2005), assuming an examinee takes a test at time t, $X_i(t)$ represents if the i^{th} item administered to the examinee has been compromised, that is,

$$X_i(t) = \begin{cases} 1, & item\ i\ is\ compromised \\ 0, & item\ i\ is\ not\ compromised \end{cases}. \tag{5}$$

Let $P(i \mid t)$ be the probability of $\{X_i(t) = 1\}$, that is, $P(i \mid t) = \Pr ob\{X_i(t) = 1\}$. Let L be the number of items in a test, $\sum_{i=1}^{L} X_i(t)$ is then the number of compromised items administered to an examinee at time *t*. Obviously, its expectation is

$$E[\sum_{i=1}^{L} X_i(t)] = \sum_{i=1}^{L} P(i \mid t), \qquad (6)$$

which is defined by Zhang and Chang as the expected number of compromised items administered to an examinee at time t. The cumulative distribution of the examinees encountering different number of compromised items was also graphed.

Results

Table 1 contains the descriptive statistics for the item parameters for the whole item pool and for the four strata, respectively. The content coverage of each stratum is similar to that of the full item pool, that is, 40% of the items are from content area one and 30% of the items are from content areas two and three, respectively. As indicated in Table 1, the distribution of *b*-parameters closely matches that of the whole item pool, and the value of the *a*-parameters increases across the strata in the stratified pool.

Based on the number of items administered at least once in the simulated CATs, Table 2 contains descriptive information on observed item exposure rates across the methods. All the 480 items are administered at least once with both the STRC-SH and randomized methods, while 332 items (69%) are used in MII-SH. MII-SH has a larger mean (0.120) and standard deviation (0.082) than those of STRC-SH (0.083; 0.043), and randomized has the best item exposure rate (0.063; 0.006). The minimum item exposure rate is 0.000 for MII-SH, 0.018 for STRC-SH, and 0.054 for randomized, while the maximum item exposure rate is 0.210 (MII-SH), 0.214 (STRC-SH), and 0.107 (randomized), respectively.

Figure 1 displays the number of items falling into different ranges of the item exposure rate. As expected, the randomized method has the best item exposure control and pool usage, while STRC-SH does not have any items that are not administered. MII-SH, on the other hand, results in a large number of items that are not used (about 31%). The SH procedure is incorporated with the STRC and MII methods; therefore, the maximum item exposure rate is approximately controlled at the prespecified 0.20 level. There are some items that exceeded the 0.20 level because of the probabilistic nature of the SH procedure.

Table 1

Descriptive Statistics for Item Parameters of the Whole Item Pool and the Four Strata

	Parameter	N	Mean	SD	Minimum	Maximum
			Item pool			
	a	480	1.056	0.347	0.193	2.685
	b	480	0.111	1.060	-2.970	2.475
	c	480	0.191	0.085	0.035	0.500
			Four strata			
1st stratum	a	120	0.766	0.220	0.193	1.550
2nd stratum		120	0.971	0.230	0.444	1.609
3rd stratum		120	1.138	0.256	0.567	1.852
4th stratum		120	1.348	0.363	0.705	2.685
1st stratum	b	120	0.111	1.058	-2.540	2.308
2nd stratum		120	0.110	1.087	-2.970	2.475
3rd stratum		120	0.111	1.051	-2.800	2.098
4th stratum		120	0.113	1.055	-2.370	2.323
1st stratum	c	120	0.195	0.077	0.040	0.500
2nd stratum		120	0.187	0.092	0.035	0.500
3rd stratum		120	0.194	0.082	0.071	0.474
4th stratum		120	0.190	0.088	0.049	0.473

Table 2

Descriptive Statistics of Observed Item Exposure Rates Across Methods

		Descriptive statistics[a]				
Method	N	Mean	SD	Minimum	Maximum
STRC-SH	480	0.083	0.043	0.018	0.214
MII-SH	332	0.120	0.082	0.000	0.210
Randomized	480	0.063	0.006	0.054	0.107

[a] Descriptive statistics are obtained based on the number of items that are administered at least once.

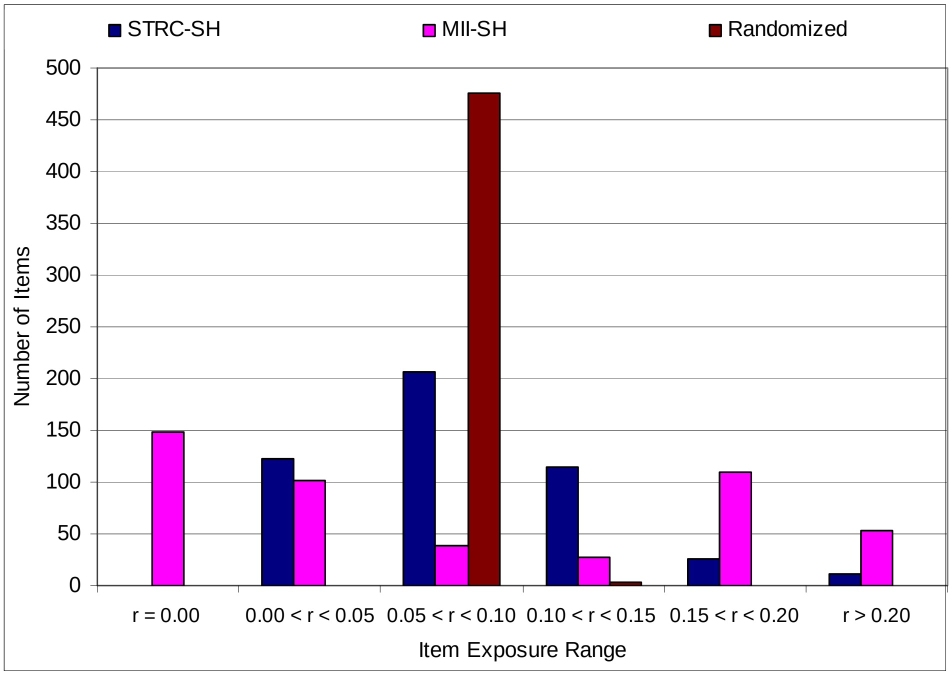


Figure 1. **Number of items falling into various ranges of observed item exposure rates (*r*) across methods.**

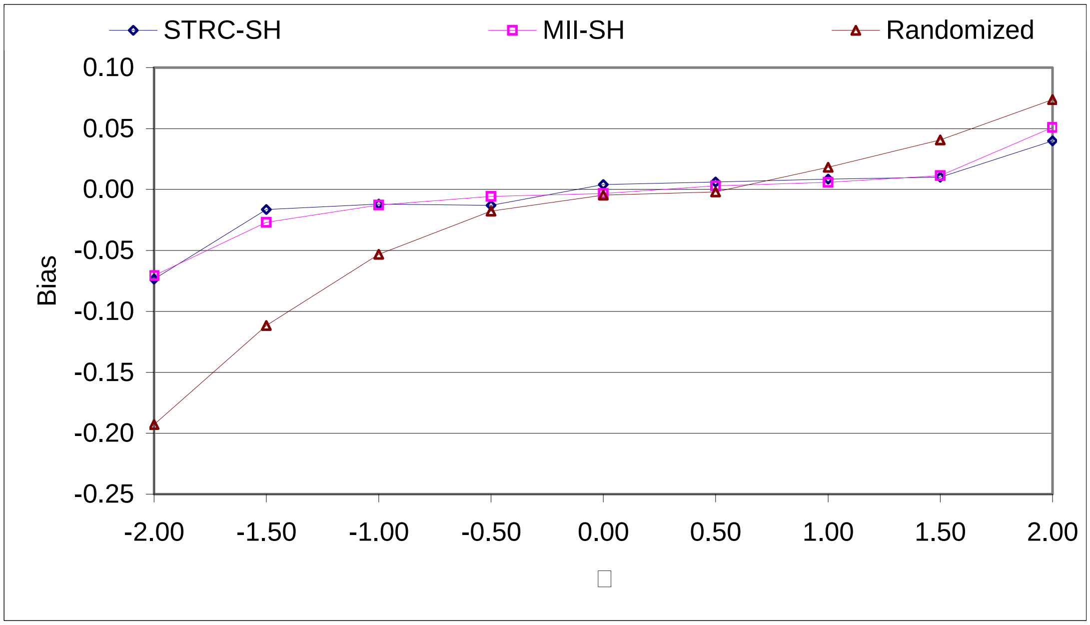
Table 3 presents the overall measurement precision and item pool usage for the CAT methods. MII-SH results in slightly better RMSE and correlation (0.226; .976) than that of STRC-SH (0.250; .971), but worse item pool usage. The difference in measurement precision is relatively small. However, the difference in item pool usage between MII-SH and STRC-SH is substantial. MII-SH results in larger χ^2 (44.007) and observed test overlap rate (17.493%) than STRC-SH (10.870 and 10.589%). The randomized method has the best item pool usage with a χ^2 of 0.082 and test overlap rate of 8.341% but the worst measurement precision (0.421; .930).

Table 3

Overall Measurement Precision and Item Pool Usage Across Methods

	Measurement precision and item pool usage				
Method	Bias	RMSE	$\rho_{\hat{\theta}_m,\theta_m}$	χ^2	Overlap
STRC-SH	-0.002	0.250	.971	10.870	10.589%
MII-SH	-0.005	0.226	.976	44.007	17.493%
Randomized	-0.029	0.421	.930	0.082	8.341%

The conditional measurement precision was computed at nine ability points from -2 to 2 in increments of 0.5 with 10,000 replications at each of the ability points. Figure 2 presents the conditional bias and RMSE, which shows that MII-SH results in slightly better conditional bias than STRC-SH but similar RMSE. The randomized method again provides the worst conditional measurement precision.



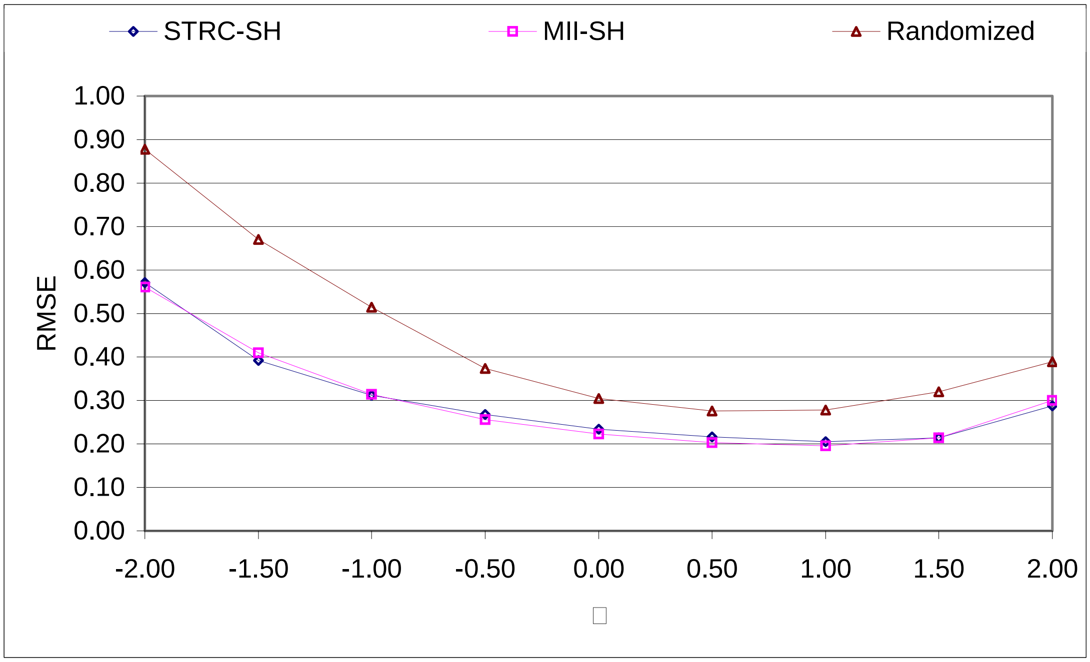


Figure 2. **Conditional bias and RMSE across methods.**

The total number of compromised items along with the descriptive statistics of the *a*-, *b*-parameters and item exposure rates of these items resulted from each condition are listed in Table 4. The number of compromised items increases as the number of thieves increases. There are fewer compromised items when MII-SH is used, but if one divides this number by the number of items that are administered, MII-SH has the largest proportion of compromised items. The difference of the number of compromised items resulting from all the conditions is small. Table 4 also lists the proportion of the item pool these compromised items comprise, which is computed based on the effect size of the pool, that is, the items that are administered at least once. According to the proportion of the item pool these compromised items comprise, STRC-SH actually results in a smaller proportion than that of MII-SH. The average item exposure rate of these compromised items for the STRC-SH method is smaller than that of MII-SH, while the randomized procedure has the smallest average item exposure rate.

Table 4

Descriptive Statistics for a-, b-Parameters, and Item Exposure Rate (r) of the Compromised Items

Methods	Condition	# of thieves	*N*	Item pool [a]	Parameters	Mean	SD	Minimum	Maximum
STRC-SH	All 10,000	10	87	0.181	*a*	1.005	0.316	0.418	1.808
					b	-0.480	1.017	-2.970	1.733
					r	0.102	0.052	0.022	0.206
MII-SH		10	77	0.232	*a*	1.096	0.359	0.416	2.447
					b	-0.428	0.923	-2.400	1.435
					r	0.167	0.057	0.008	0.207
Randomized		10	94	0.196	*a*	1.049	0.363	0.193	1.794
					b	0.236	1.067	-2.540	2.308
					r	0.062	0.002	0.056	0.067
STRC-SH	All 10,000	20	151	0.315	*a*	1.019	0.350	0.418	2.685
					b	-0.264	1.090	-2.970	1.733
					r	0.099	0.051	0.021	0.206
MII-SH		20	132	0.398	*a*	1.160	0.384	0.416	2.685
					b	-0.181	1.028	-2.400	1.733
					r	0.167	0.055	0.008	0.208

(Table continues)

Table 4 (continued)

Method	Condition	# of thieves	*N*	Item pool [a]	Parameters	Mean	SD	Minimum	Maximum
Randomized		20	169	0.352	*a*	1.049	0.354	0.193	1.852
					b	0.105	1.098	-2.970	2.308
					r	0.062	0.002	0.056	0.067
STRC-SH	All 10,000	30	206	0.429	*a*	1.036	0.332	0.418	2.685
					b	-0.163	1.020	-2.970	1.733
					r	0.099	0.049	0.021	0.214
MII-SH		30	163	0.491	*a*	1.196	0.373	0.416	2.685
					b	-0.086	0.975	-2.400	1.733
					r	0.170	0.052	0.008	0.209
Randomized		30	225	0.469	*a*	1.055	0.347	0.193	2.107
					b	0.161	1.073	-2.970	2.308
					r	0.062	0.002	0.056	0.067
STRC-SH	First 1,000	10	91	0.190	*a*	1.001	0.316	0.418	1.808
					b	-0.491	0.849	-2.800	1.170
					r	0.110	0.051	0.022	0.214
MII-SH		10	81	0.244	*a*	1.149	0.342	0.515	2.447
					b	-0.303	0.821	-2.400	1.435
					r	0.176	0.049	0.008	0.207
Randomized		10	96	0.200	*a*	1.016	0.327	0.487	1.790
					b	0.183	1.032	-2.970	2.323
					r	0.062	0.002	0.055	0.066
STRC-SH	First 1,000	20	159	0.331	*a*	1.000	0.347	0.193	2.685
					b	-0.370	1.035	-2.970	2.107
					r	0.103	0.052	0.018	0.214
MII-SH		20	133	0.401	*a*	1.155	0.350	0.416	2.478
					b	-0.185	0.968	-2.400	1.903
					r	0.170	0.052	0.008	0.209
Randomized		20	165	0.344	*a*	1.063	0.337	0.418	2.135
					b	0.303	1.052	-2.970	2.323
					r	0.062	0.002	0.055	0.069
STRC-SH	First 1,000	30	208	0.433	*a*	1.015	0.332	0.193	2.685
					b	-0.265	0.966	-2.970	2.107
					r	0.100	0.048	0.018	0.214

(Table continues)

Table 4 (continued)

Method	Condition	# of thieves	N	Item pool [a]	Parameters	Mean	SD	Minimum	Maximum
MII-SH		30	170	0.512	*a*	1.180	0.350	0.416	2.685
					b	-0.087	0.947	-2.400	1.903
					r	0.169	0.051	0.008	0.209
Randomized		30	218	0.454	*a*	1.059	0.336	0.375	2.135
					b	0.270	1.096	-2.970	2.475
					r	0.062	0.002	0.055	0.070
STRC-SH	First 5,000	10	85	0.177	*a*	1.005	0.330	0.418	2.107
					b	-0.455	1.105	-2.970	1.473
					r	0.104	0.055	0.022	0.206
MII-SH		10	83	0.250	*a*	1.098	0.358	0.416	2.447
					b	-0.347	0.992	-2.400	1.480
					r	0.169	0.053	0.008	0.207
Randomized		10	93	0.194	*a*	1.020	0.359	0.487	1.852
					b	0.071	1.114	-2.970	2.308
					r	0.062	0.002	0.057	0.067
STRC-SH	First 5,000	20	160	0.333	*a*	1.044	0.339	0.405	2.685
					b	-0.214	1.021	-2.970	1.620
					r	0.105	0.051	0.022	0.214
MII-SH		20	143	0.431	*a*	1.187	0.361	0.416	2.685
					b	-0.086	0.978	-2.970	1.733
					r	0.173	0.049	0.008	0.209
Randomized		20	164	0.342	*a*	1.024	0.339	0.193	1.852
					b	0.086	1.065	-2.970	2.308
					r	0.062	0.002	0.056	0.067
STRC-SH	First 5,000	30	214	0.446	*a*	1.063	0.345	0.405	2.685
					b	-0.060	1.029	-2.970	2.323
					r	0.101	0.050	0.022	0.214
MII-SH		30	173	0.521	*a*	1.197	0.361	0.416	2.685
					b	-0.015	0.984	-2.970	1.896
					r	0.168	0.055	0.008	0.209
Randomized		30	228	0.475	*a*	1.043	0.346	0.193	2.685
					b	0.093	1.039	-2.970	2.323
					r	0.062	0.002	0.055	0.069

(Table continues)

Table 4 (continued)

Method	Condition	# of thieves	*N*	Item pool [a]	Parameters	Mean	SD	Minimum	Maximum
STRC-SH	Last 5,000	10	87	0.181	*a*	1.121	0.350	0.416	2.685
					b	0.368	0.990	-1.710	2.323
					r	0.096	0.041	0.032	0.203
		10	89	0.268	*a*	1.266	0.342	0.793	2.685
MII-SH					*b*	0.376	0.944	-1.970	2.308
					r	0.166	0.054	0.006	0.209
		10	88	0.183	*a*	1.065	0.328	0.193	1.748
Randomized					*b*	0.091	0.986	-2.970	1.944
					r	0.062	0.002	0.056	0.067
	Last 5,000	20	156	0.325	*a*	1.076	0.335	0.416	2.685
STRC-SH					*b*	0.050	0.932	-2.420	2.323
					r	0.101	0.046	0.018	0.214
MII-SH		20	142	0.428	*a*	1.200	0.323	0.649	2.685
					b	0.096	0.963	-2.070	2.308
					r	0.166	0.055	0.006	0.209
Randomized		20	163	0.340	*a*	1.073	0.337	0.193	1.852
					b	0.103	0.995	-2.970	1.944
					r	0.062	0.002	0.055	0.067
STRC-SH	Last 5,000	30	218	0.454	*a*	1.110	0.338	0.416	2.685
					b	0.201	0.921	-2.420	2.323
					r	0.097	0.045	0.018	0.214
MII-SH		30	170	0.512	*a*	1.196	0.332	0.649	2.685
					b	0.178	0.994	-2.070	2.308
					r	0.158	0.063	0.000	0.209
Randomized		30	214	0.446	*a*	1.047	0.341	0.193	2.107
					b	0.094	1.039	-2.970	2.475
					r	0.062	0.002	0.055	0.067

[a] Proportion of item pool is computed based on the number of items that have been administered at least once.

Table 5 contains the average number of compromised items encountered by examinees under each condition. The average number of compromised items increases as the number of thieves increases. MII-SH results in a higher average number of compromised items than that of

STRC-SH, while the randomized method has the smallest average number of compromised items. The difference between the average number of compromised items resulting from the randomized and STRC-SH methods is about two, and this difference can be as high as eight between STRC-SH and MII-SH. The damage caused by organized item theft is the most severe when the thieves appear in the first 1,000 examinees. The average number of compromised items examinees encounter can be as high as about 29 with the MII-SH method when there are 30 thieves coming from the first 1,000 examinees, and this number is about 14 when there are 10 thieves. The damage is even more severe when the thieves come from the first 5,000 examinees rather than from all 10,000 examinees, while the damage is the smallest if the thieves are from the last 5,000 examinees. The results presented in Table 5 indicate that if organized item theft happens at the beginning of a CAT, the damage caused by such an act will be the most severe because examinees who take the test later would potentially get benefits by studying these compromised items.

Table 5

Average Number of Compromised Items Each Examinee Encountered Under Different Conditions

Method	Condition	Number of thieves	Number of compromised items
STRC-SH	All 10,000	10	5.803
MII-SH			8.675
Randomized			4.889
STRC-SH		20	10.277
MII-SH			16.102
Randomized			9.125
STRC-SH		30	14.186
MII-SH			21.318
Randomized			12.172
STRC-SH	First 1,000	10	9.705
MII-SH			13.807
Randomized			7.660

(Table continues)

Table 5 (continued)

Method	Condition	Number of thieves	Number of compromised items
STRC-SH		20	15.726
MII-SH			21.896
Randomized			13.125
STRC-SH		30	20.065
MII-SH			27.821
Randomized			17.318
STRC-SH	First 5,000	10	7.906
MII-SH			12.715
Randomized			6.812
STRC-SH		20	14.447
MII-SH			21.525
Randomized			11.626
STRC-SH		30	18.705
MII-SH			25.574
Randomized			16.267
STRC-SH	Last 5,000	10	3.260
MII-SH			5.933
Randomized			2.836
STRC-SH		20	5.668
MII-SH			8.865
Randomized			4.769
STRC-SH		30	7.663
MII-SH			10.666
Randomized			6.306

Figures 3 to 5 display the cumulative percentage of compromised items calculated from the 10,000 simulated examinees under different simulation conditions. The damage caused by organized item theft in terms of the cumulative percentage is affected by the sequence of thieves' appearances (e.g., thieves are randomly selected from the first 1,000 examinees), the number of thieves, and the CAT designs. The damage is the most severe if the thieves appear at the early stage of the testing whereas it is the least severe if they appear at the later stage. Among the three CAT designs, the MII-SH method was affected the most by the simulated item theft activity. As expected, the damage was more severe if there were more thieves. As shown in Figure 5, about 20% of the examinees encountered fewer than 11 compromised items with the STRC-SH method if 30 thieves appeared in the first 1,000 examinees, while about 21% of the examinees encountered fewer than 24 compromised items with MII-SH.

Discussion

Organized item theft can cause serious damage to a large-scale CAT program that has high-stakes outcomes. Chang and Zhang (2002, 2003) initiated a theoretical basis for examining the severity of possible test security violations of organized item theft in large-scale computerized testing. Based on the randomized item selection assumption, Chang and Zhang derived indices for evaluating the severity of possible test security violations, which can serve as lower bounds for test overlap rates. The theoretical results may help practitioners to design more secure computerized tests. A test security panel may evaluate the discrepancy between the theoretical lower bound and the observed test overlap rate generated by the item selection algorithm under investigation. A large discrepancy indicates that the algorithm needs to be further improved by lowering the observed test overlap rate, and a small difference shows that the item selection algorithm generates satisfactory results.

According to Chang and Zhang (2002, 2003), the damage caused by organized item theft can be lessened by increasing item pool size and selecting items more evenly. In practice, however, the randomized item selection procedure is never used due to its poor measurement precision. The goal of the current research was to assess the severity caused by organized item theft using two more realistic CAT designs while employing the randomized item selection method as a baseline for comparison. The results of the simulation study indicated that the damage increases as the number of thieves increases. The degree of the damage is also related to the time when the item theft takes place. If the thieves were administered tests earlier, the damage was more severe than when they took tests later. The damage would be severe even if the thieves were randomly appointed among the 10,000 simulees. However, the damage became less severe when item theft occurred after 50% of the total numbers of examinees had taken the test. The conclusion held for all three item selection methods regardless how many thieves were involved.

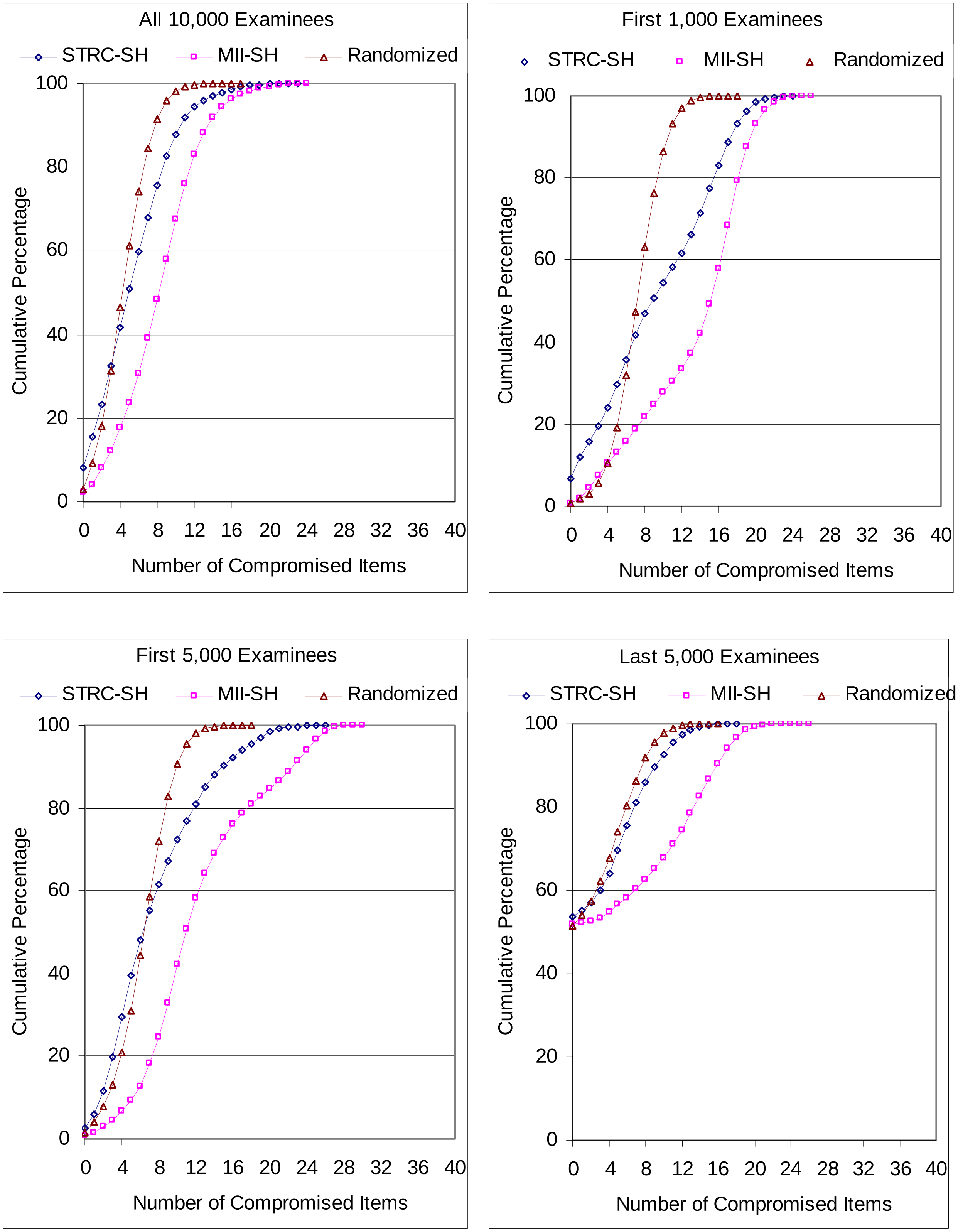


Figure 3. **Cumulative percentage of number of compromised items with 10 thieves randomly selected.**

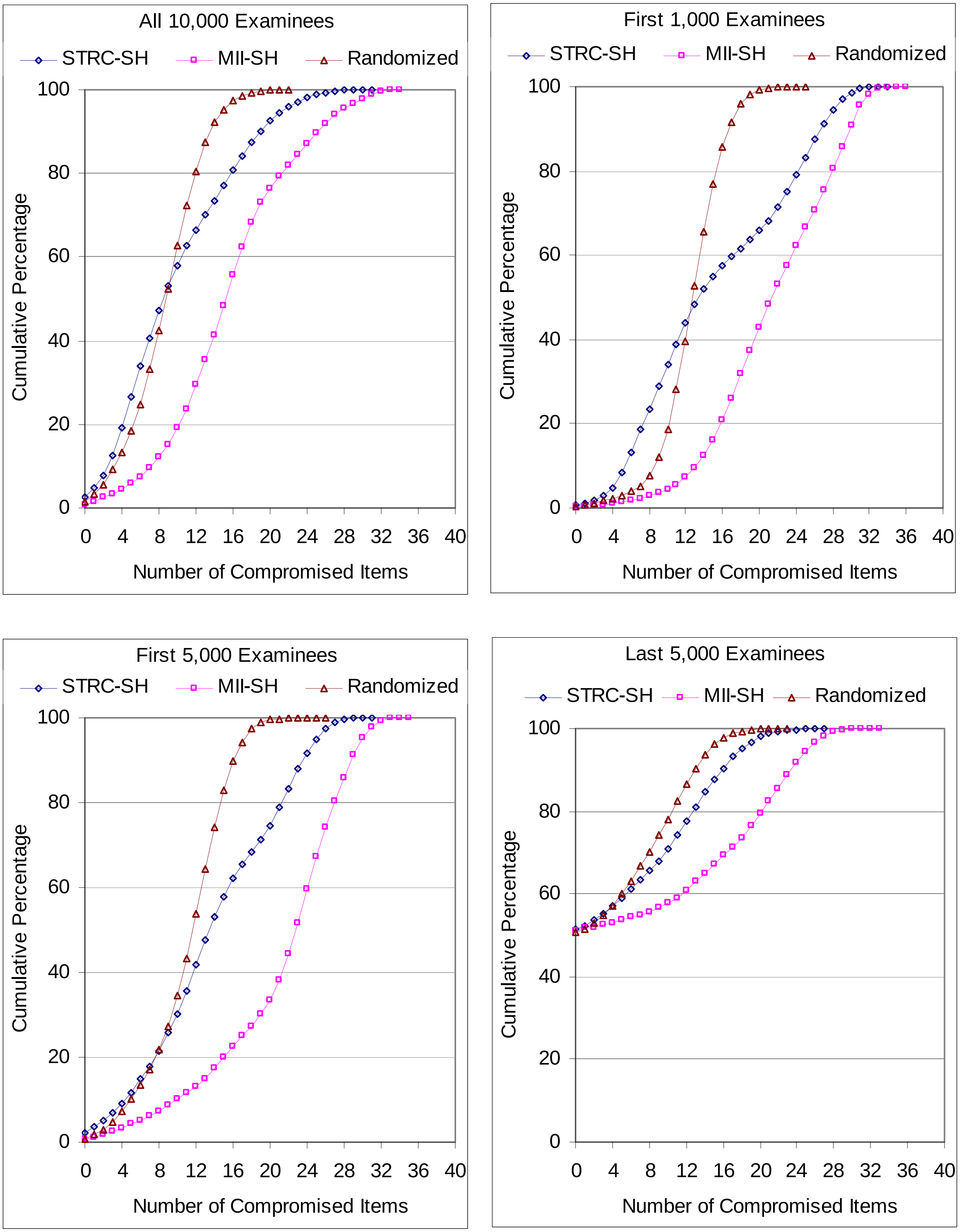


Figure 4. **Cumulative percentage of number of compromised items with 20 thieves randomly selected.**

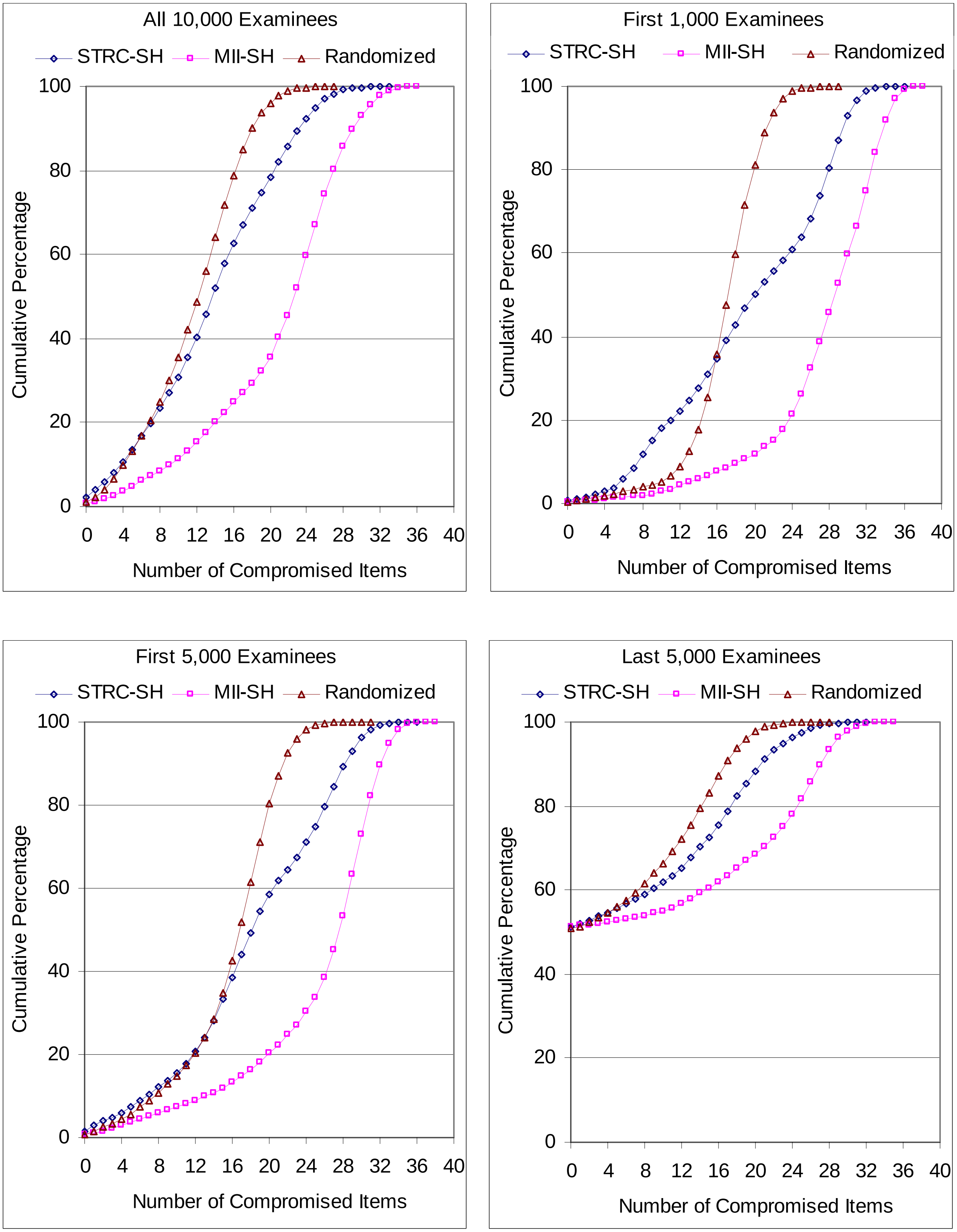


Figure 5. **Cumulative percentage of number of compromised items with 30 thieves randomly selected.**

Chang and Zhang (2002, 2003) proposed to use the expected number of compromised items as a vital test security index. However, our simulation results revealed that although the total number of compromised items resulting from MII-SH is smaller than the numbers resulting from both STRC-SH and the randomized method, the disadvantages of using MII-SH are much higher. When MII-SH is used, the number of compromised items examinees encounter is much higher than with STRC-SH and the randomized methods. The reason is that the number of unadministered items using MII-SH is larger than either STRC-SH or the randomized methods, which makes the actual effective item pool size for MII-SH much smaller. MII-SH yields the highest observed average test overlap rate, and the likelihood that an examinee will access compromised items is much higher, which explains in the simulation results why examinees could come across more compromised items. The STRC-SH method, on the other hand, results in much better test security control because it has a better item pool usage and smaller test overlap rate. Clearly, the potential damage caused by organized item theft is less severe.

The effect size of an item pool has played an essential role in our investigation. It is well-known that maximum information-based item selection methods tend to select the optimal items more often and typically leave a large proportion of an item pool remain unused. In the current study, the MII-SH method used 69% of the items and the remaining 31% were never administered, thus the usable pool size turned out to be 332 rather than 480 items. As a consequence, the observed average test overlap rate was much higher for MII-SH than that of STRC-SH (17.493% versus 10.589%).

With the smallest test overlap rate, the randomized method can be considered to have the best test security control. However, STRC-SH performs quite similarly to that of the randomized method in terms of test security. As for measurement precision, the performance of STRC-SH is very close to MII-SH. Therefore, the overall performance of STRC-SH can be considered the best of the three methods employed in the current simulation.

Finally, based on only two CAT item selection methods, the current study is an initial attempt to empirically investigate the severity of possible test security violations in CAT. More studies along this line of research are needed. The results clearly indicated that item selection methods being used in operational CAT could and should be evaluated by this kind of simulation study. A test security panel can then evaluate the observed severity indices generated by the item selection algorithm under investigation. In future studies, more issues should be examined, for example, the effects of item pool size on the severity of possible test security violations, the influence of thieves’ memory capacities on security, the effectiveness of using a multiple item pool approach, test security issue may be evaluated conditionally with respect to ability, and different thievery models, such as based on the efforts thieves may be spending on stealing items according to the stakes of an exam, can be developed.

References

Chang, H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement, 25*(4), 333–341.

Chang, H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222.

Chang, H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67,* 387–398.

Chang, H., & Zhang, J. (2003, April). *Assessing CAT security breaches by the item pooling index.* Paper presented at the annual meeting of National Council on Measurement in Education, Chicago, IL.

Chen, S., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145.

Davey, T., & Parshall, C. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Hetter, R., & Sympson, B. (1997). Item exposure control in CAT-ASVAB. In W. Sands, B. Waters, & J. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp.141–144). Washington, DC: American Psychological Association.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R., & Bock, R. D. (1982). BILOG: Item analysis and test scoring with binary logistic models [Computer software]. Mooresville, IN: Scientific Software, Inc.

Stocking, M., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23,* 57-75.

Sympson, J., & Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

Thomasson, G. (1995, June). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis, MN.

Way, W. D. (1998). Protect the integrity of computerizes testing item pools. *Educational Measurement: Issues and Practice*, *Winter,* 17–27.

Yi, Q., & Chang, H. (2003). *a*-Stratified multistage CAT design with content-blocking. *British Journal of Mathematical and Statistical Psychology, 56*, 359–378.

Zhang, J., & Chang, H. (2005). *The effectiveness of enhancing test security by using multiple item pools* (ETS RR-05-19). Princeton, NJ: ETS.